HOUSING DATA ANALYSIS

FOR HOMES IN KING COUNTY, WA

Dennis Trimarchi

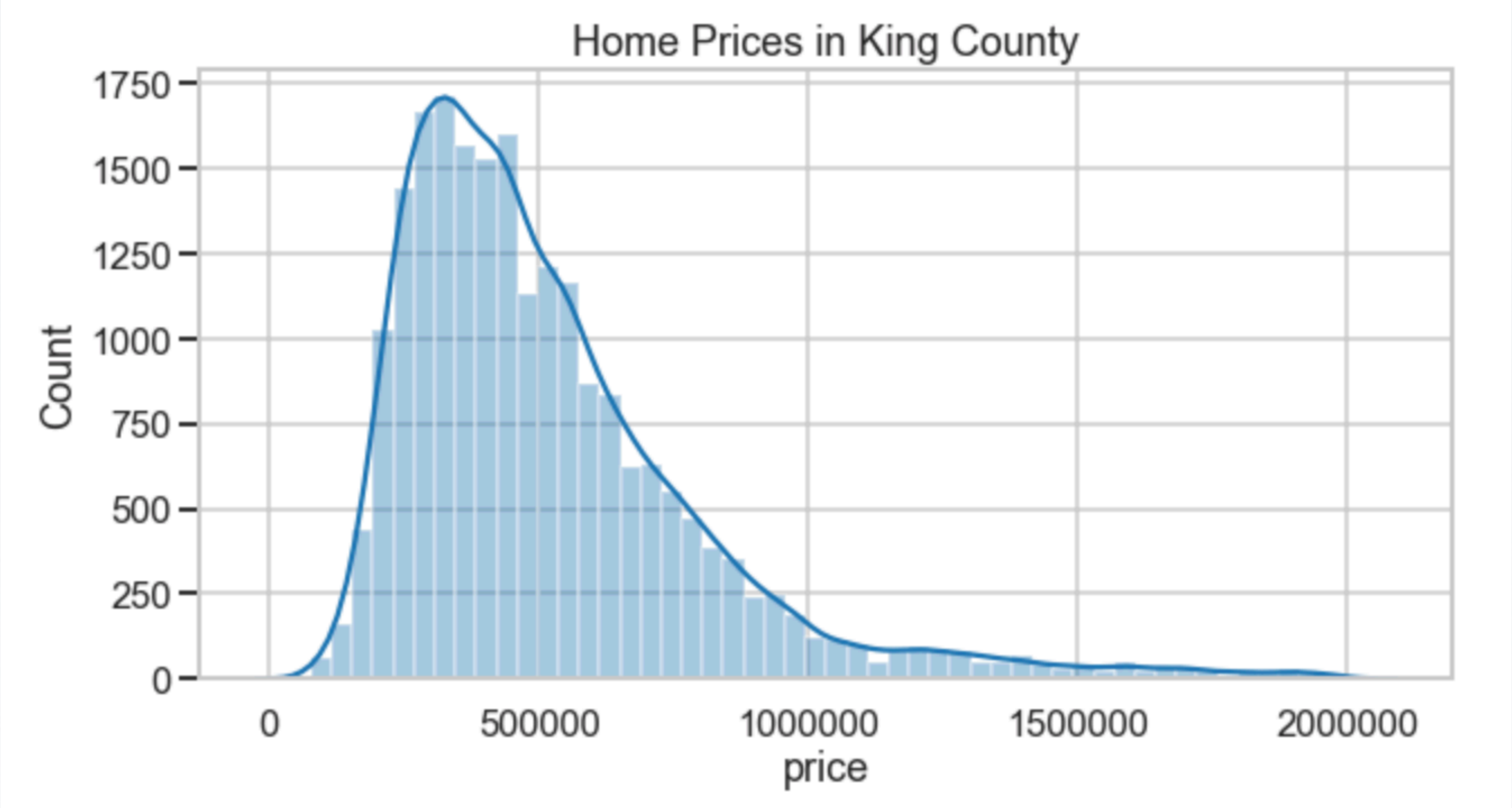
2019-06-02

MOTIVATION & INITIAL ANALYSIS

mean:
~ **$520K**

median:
~**$450K**

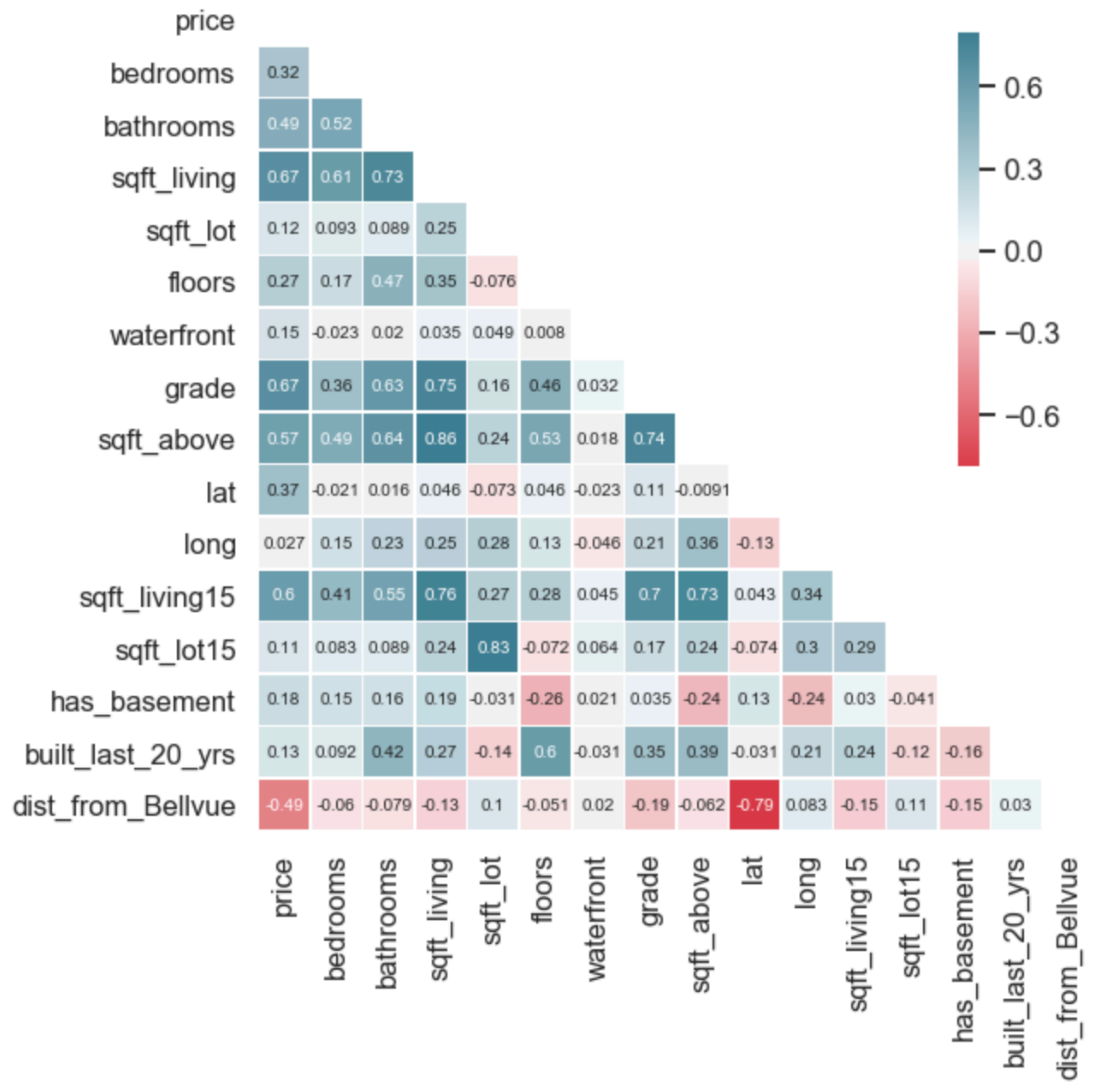
std dev:
~**$286K**



- **Making predictions is useful**! Assessors make a living off of predicting home values. We want to see if an algorithm can be developed to accurately predict a home's price in King County, WA.

- A reliable set of data is needed: Removed 634 out of 21597 records that had missing / spurious data. **Removed homes valued over $2M** because they are not comparable to the majority of homes. Even at the $2M cutoff, the distribution is still skewed high.

CORRELATION MATRIX



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	grade	sqft_above	lat	long	sqft_living15	sqft_lot15	has_basement	built_last_20_yrs	dist_from_Bellvue
price																
bedrooms	0.32															
bathrooms	0.49	0.52														
sqft_living	0.67	0.61	0.73													
sqft_lot	0.12	0.093	0.089	0.25												
floors	0.27	0.17	0.47	0.35	-0.076											
waterfront	0.15	-0.023	0.02	0.035	0.049	0.008										
grade	0.67	0.36	0.63	0.75	0.16	0.46	0.032									
sqft_above	0.57	0.49	0.64	0.86	0.24	0.53	0.018	0.74								
lat	0.37	-0.021	0.016	0.046	-0.073	0.046	-0.023	0.11	-0.0091							
long	0.027	0.15	0.23	0.25	0.28	0.13	-0.046	0.21	0.36	-0.13						
sqft_living15	0.6	0.41	0.55	0.76	0.27	0.28	0.045	0.7	0.73	0.043	0.34					
sqft_lot15	0.11	0.083	0.089	0.24	0.83	-0.072	0.064	0.17	0.24	-0.074	0.3	0.29				
has_basement	0.18	0.15	0.16	0.19	-0.031	-0.26	0.021	0.035	-0.24	0.13	-0.24	0.03	-0.041			
built_last_20_yrs	0.13	0.092	0.42	0.27	-0.14	0.6	-0.031	0.35	0.39	-0.031	0.21	0.24	-0.12	-0.16		
dist_from_Bellvue	-0.49	-0.06	-0.079	-0.13	0.1	-0.051	0.02	-0.19	-0.062	-0.79	0.083	-0.15	0.11	-0.15	0.03	

Price has strongest positive correlation with:

sqft_living,
grade,
sqft_living15,
sqft_above,
bathrooms

Price has strongest negative correlation with:

dist_from_Bellvue

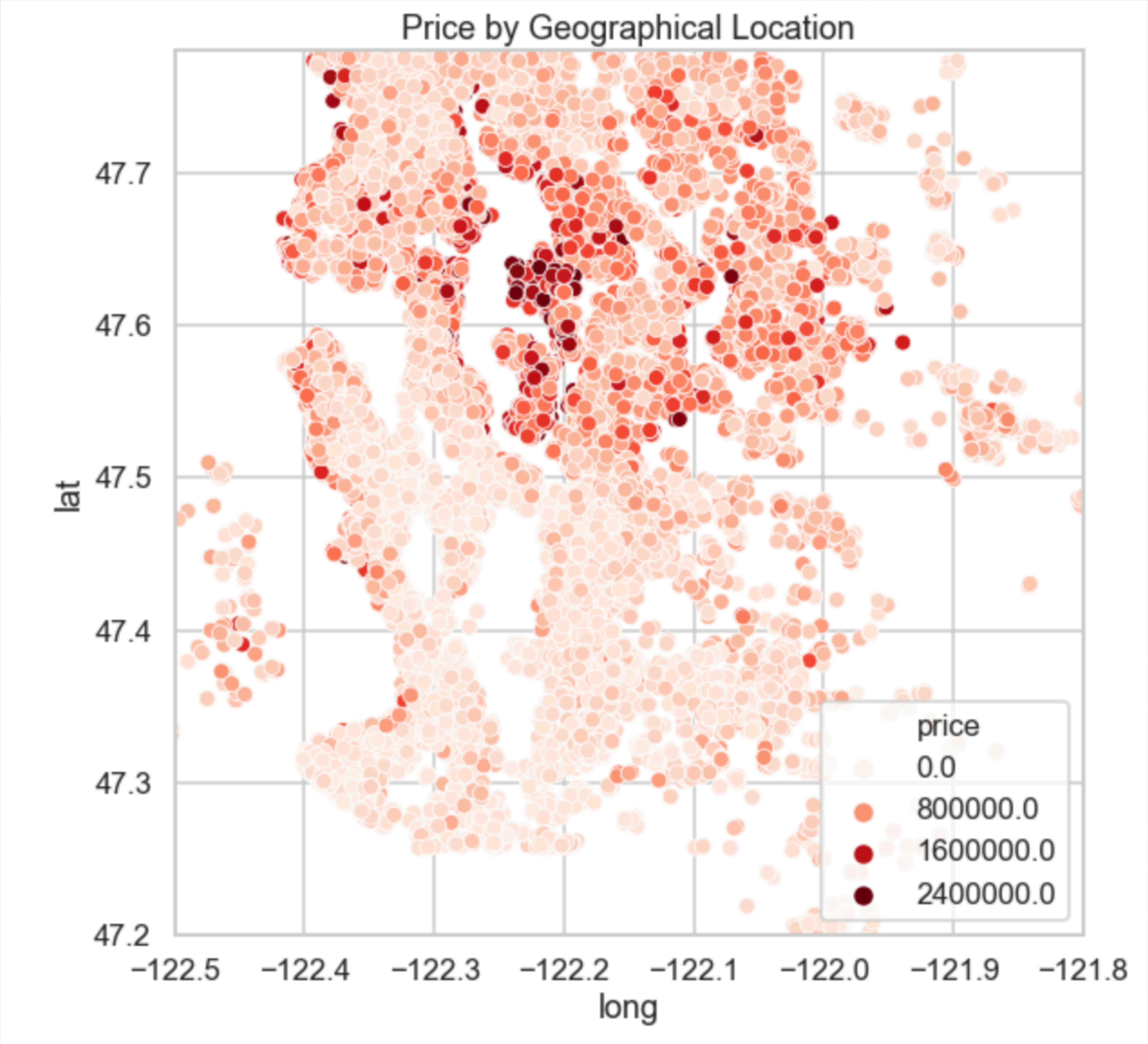
Many of the features are correlated with each other in addition to price. This is bad model development as independent variables should be independent.

GEOGRAPHY

Some expensive homes in King County are clustered in a general area.

The darkest area of the map, and the center of a region of more expensive homes is Bellevue, WA.

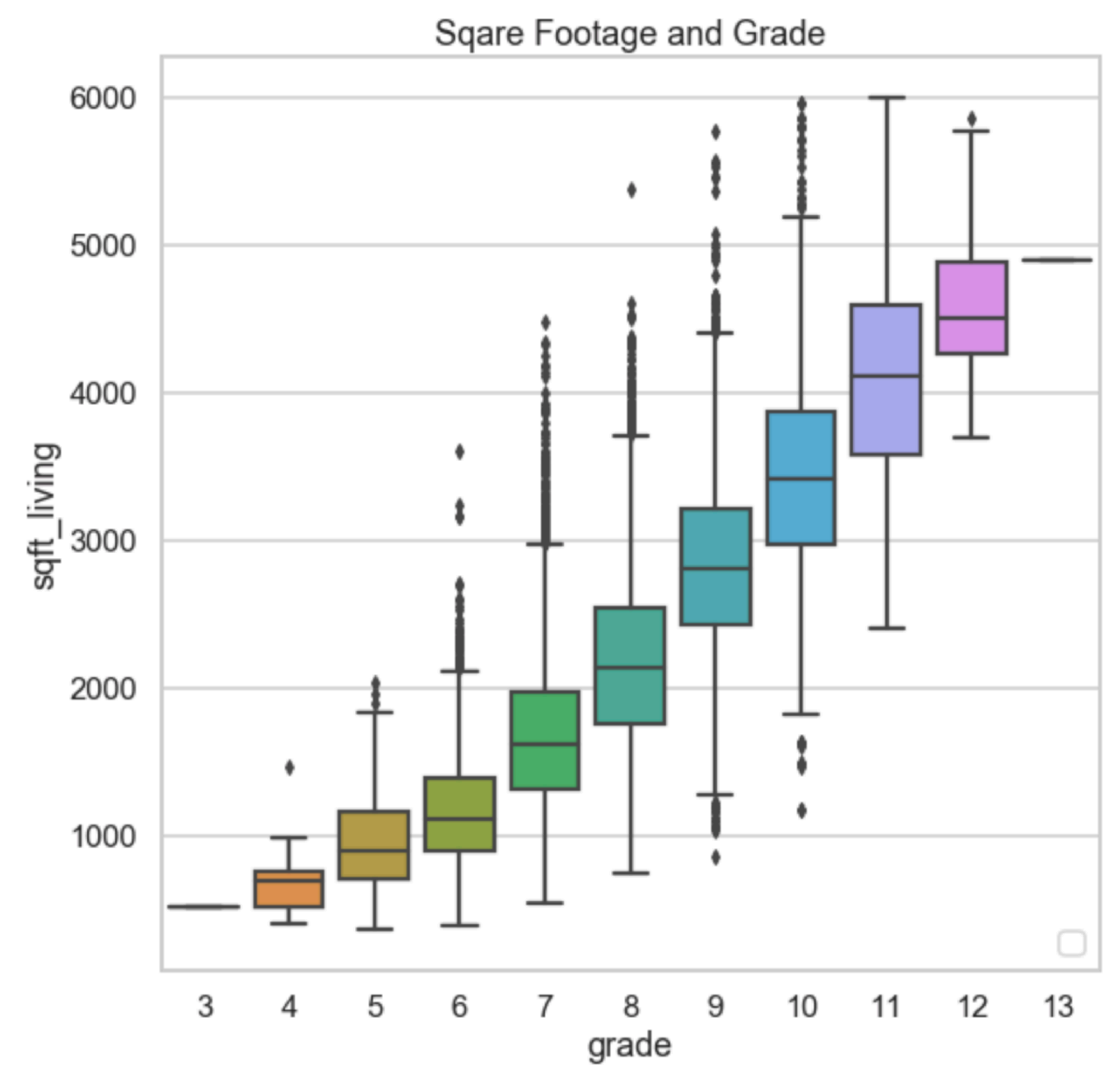
Having a home in this region will impact price.



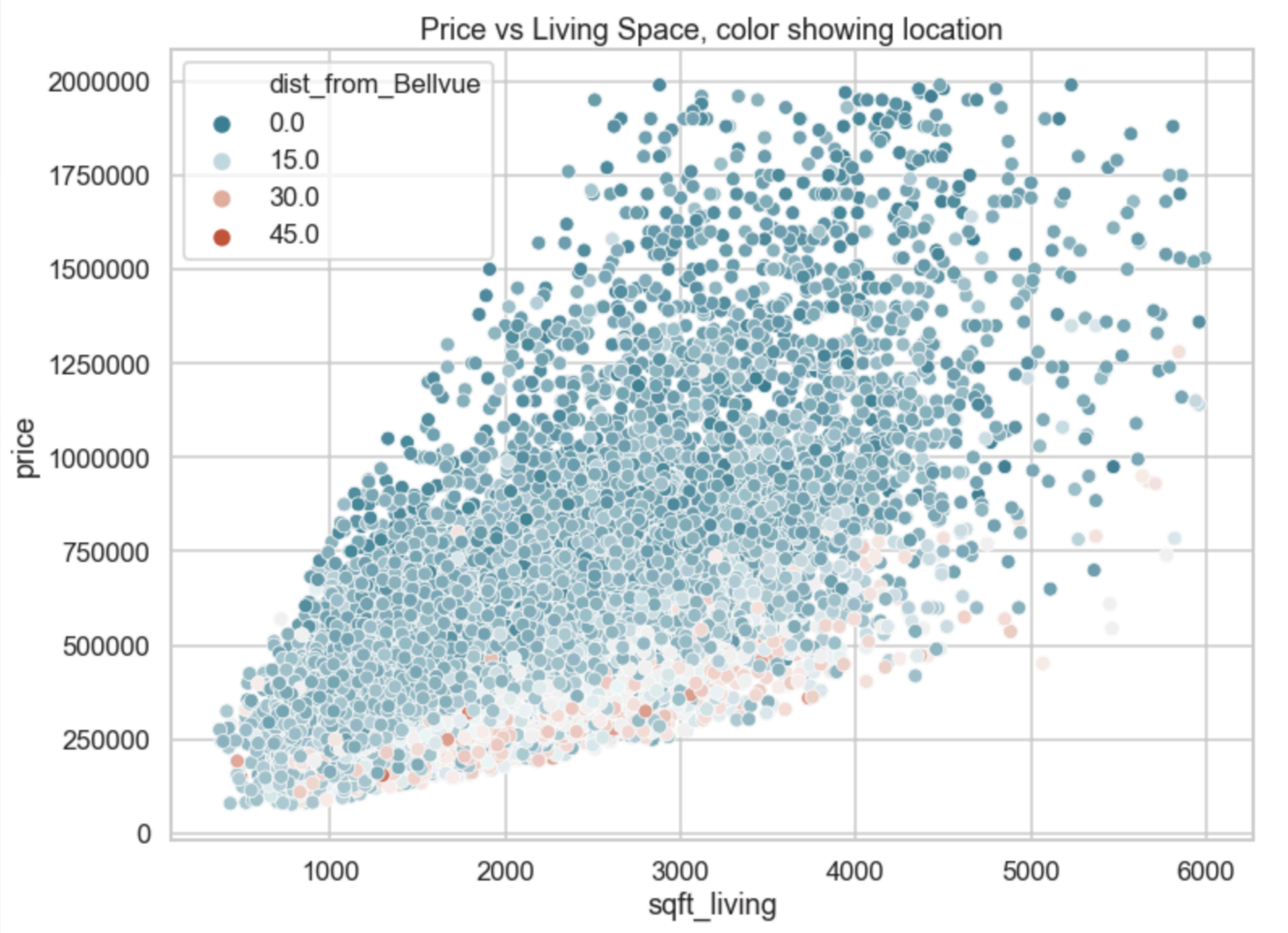
SQ FT LIVING & GRADE

These two fields are so highly correlated that you could almost use one to predict the other. —Higher grades are given to larger homes. This is supported by the high correlation coefficient between the two features.

grade / average price	
3	$262K
4	$212K
5	$247K
6	$301K
7	$401K
8	$540K
9	$762K
10	$993K
11	$1,209K
12	$1,541K
13	$1,780K



LIVING SPACE VS PRICE



Clear positive relationship between **price and living space**.

Also, distance from Bellevue, WA shown through the use of color.

As you can see, the further distances are clustered at the bottom of the distribution indicating lower price for being farther away.

ALGORITHM & DISCUSSION

Modeled Price as a function of sqft_living and dist_from_Bellvue.These particular features were selected based on high correlation to price and low correlation with each other.This makes for a stronger model.

$$price = 202.8 * sqft\_living \; - \; 196{,}800 * \ln(dist\_from\_Bellvue) \; + \; 550{,}300$$

The Model is simple in that it only contains two features and it uses information that should be readily available for any home in King County: square footage, and location.

There are some limitations in this model. Firstly, it can only be applied to data for King County. Homes across the state in other counties could not realistically use distance from Bellevue,WA as a way to predict home price.

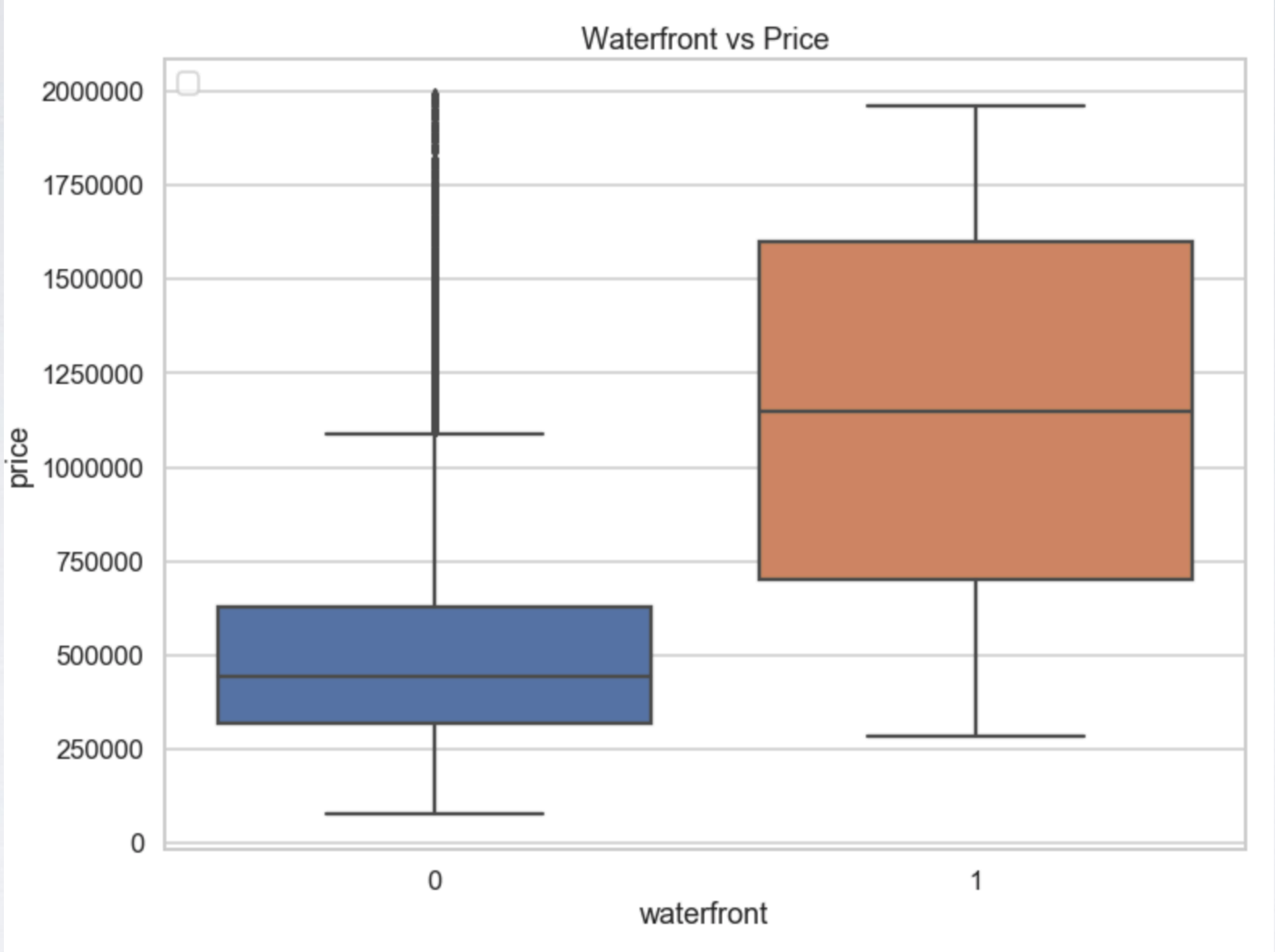
Secondly, the algorithm only explains 61% of the variation in home price for the dataset. Other factors come into play when it comes to price.

FURTHER WORK

- Some additional information gathering would improve this model.
- Waterfront: In the Seattle area waterfront property is everywhere, and it would be nice to see if having a waterfront view impacts price. Unfortunately, for many of the homes this information was not provided, as such this feature was not included in the model.
- Sqft_living vs Grade was chosen over grade even though both features are highly correlated with price. With more data, it's quite possible that grade could be a better predictor. Also, it would be beneficial to know how the details of the King County grading system works.
- Lot Size: Another feature that could be investigated further is lot size. The data seemed to indicate a - city vs. rural split.

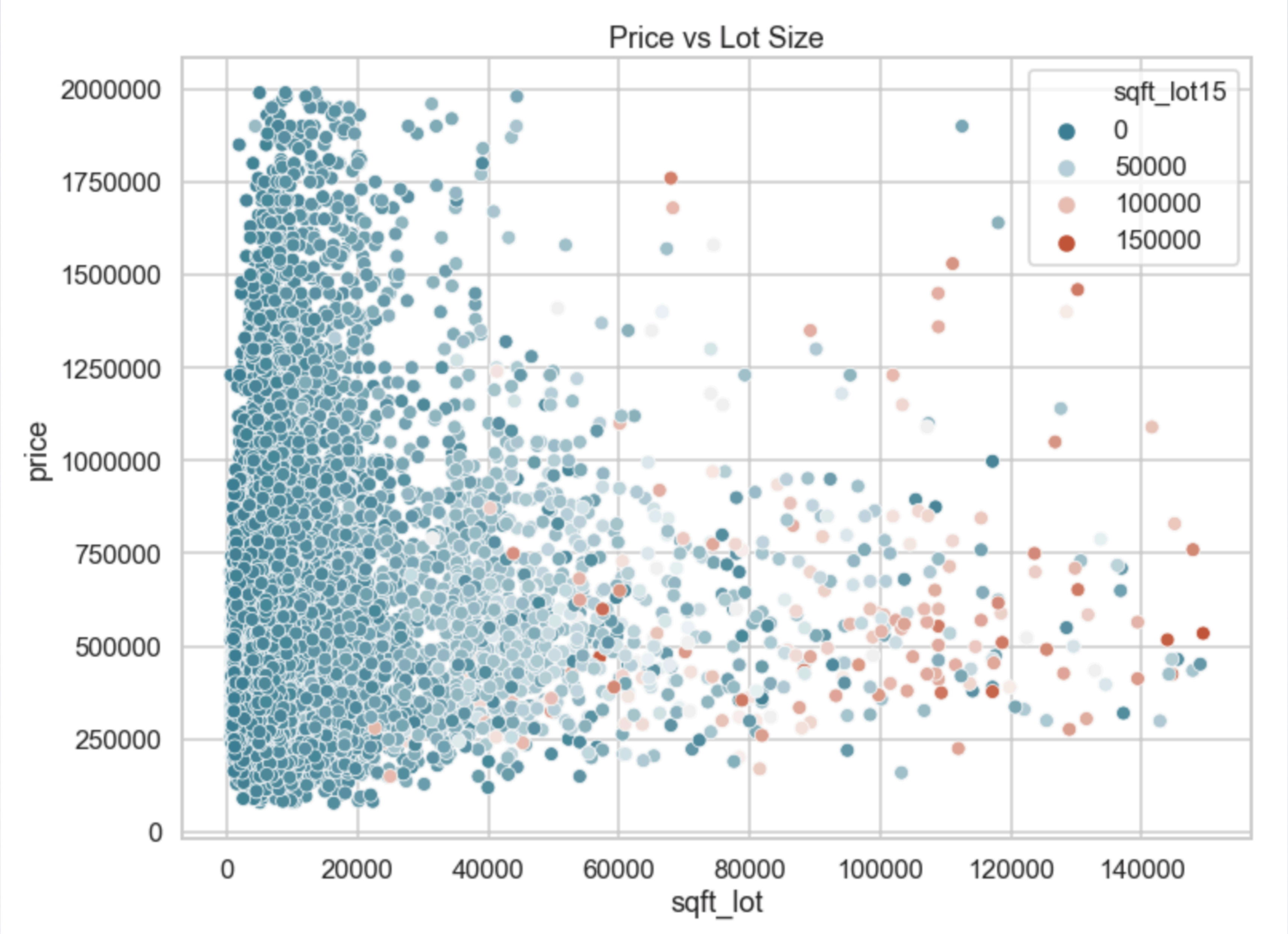
BACKUP SLIDES

WATERFRONT



- Clear difference in price for waterfront properties.

LOT SIZE



- Having neighbors with large lot sizes (sqft_lot15) could indicate more rural areas.